Exploiting SPATIAL CONSISTENCY is important, and BLURS BEHAVE LIKE ENSEMBLES

NAMUK PARK, SONGKUK KIM
NAVER AI Lab, Yonsei University

Blurs Behave Like Ensembles

We introduce a novel ensemble method, ***"spatial ensemble"***. Spatial ensemble improve accuracy, uncertainty, and robustness *without* increasing inference time. But what is a spatial ensemble? Spatial ensemble, or spatial smoothing, is a method that aggregate nearby feature maps (See ⇘).

I. How Can We Apply Spatial Smoothing to NNs?

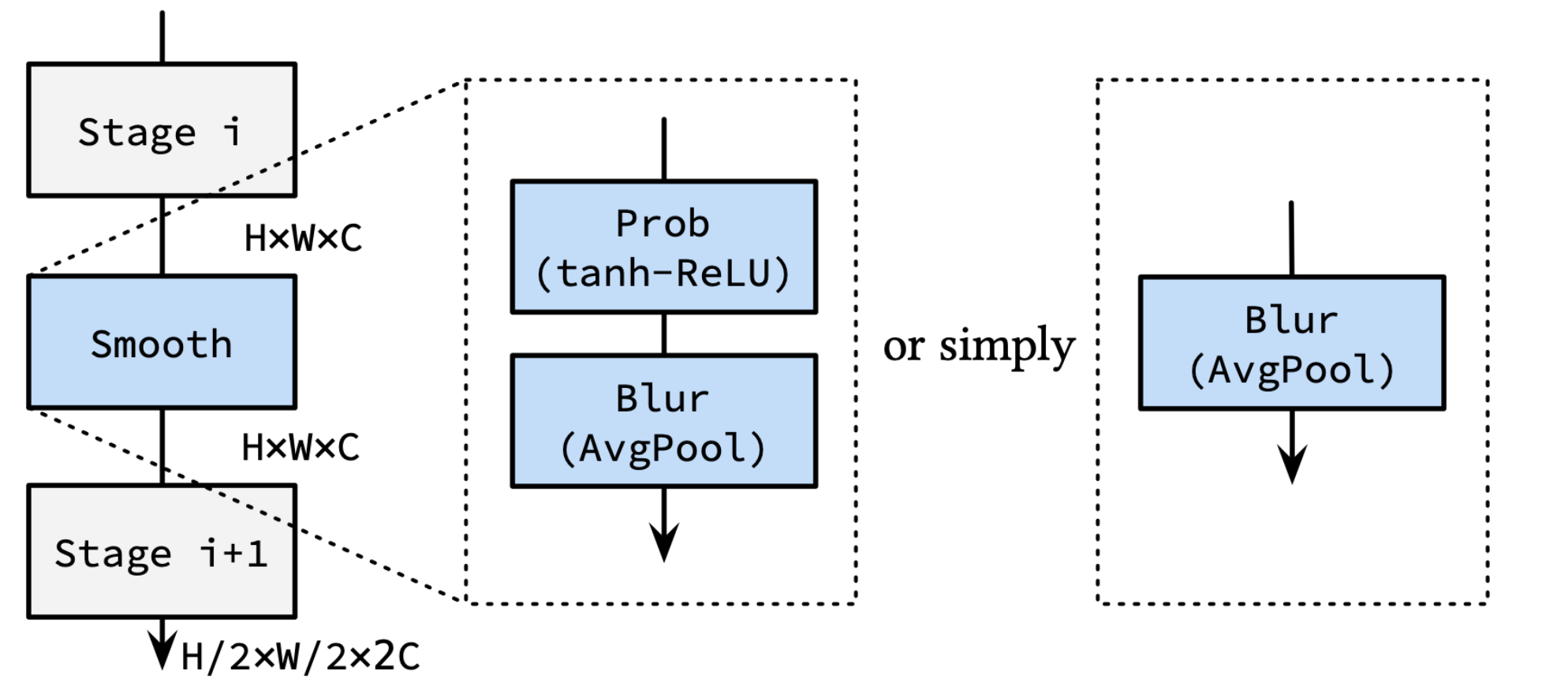


Figure 1. Simply add

```
AvgPool2d(kernel_size=2,
stride=1, padding=1)
```

at the end of stages. For example, use four spatial smoothing layers for ResNet.

II. Spatial Smoothing Improves MC Dropout

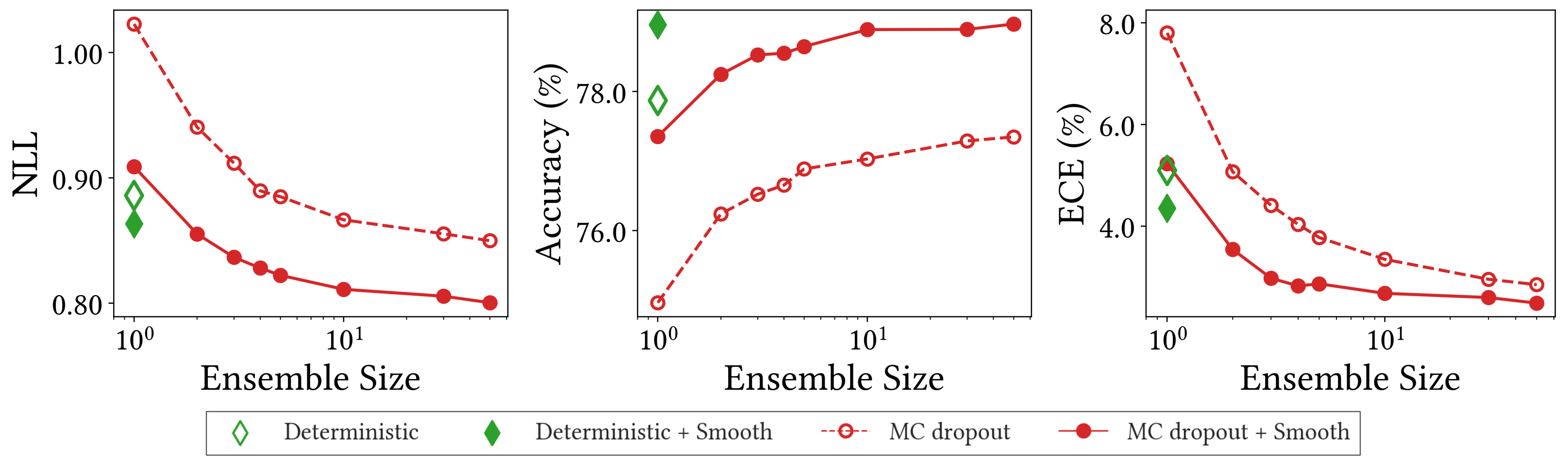
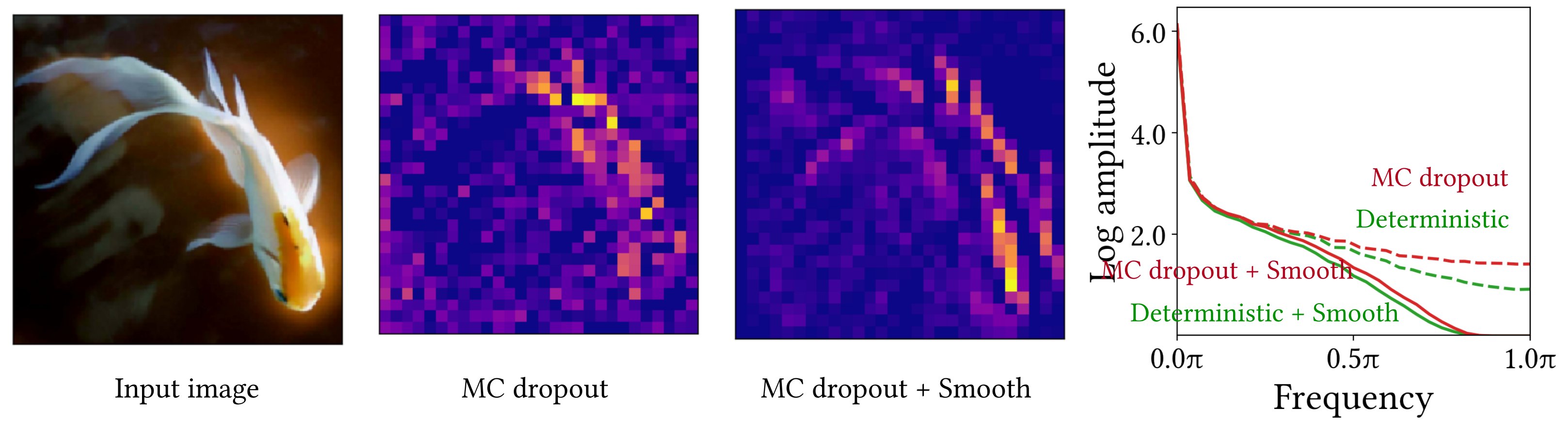


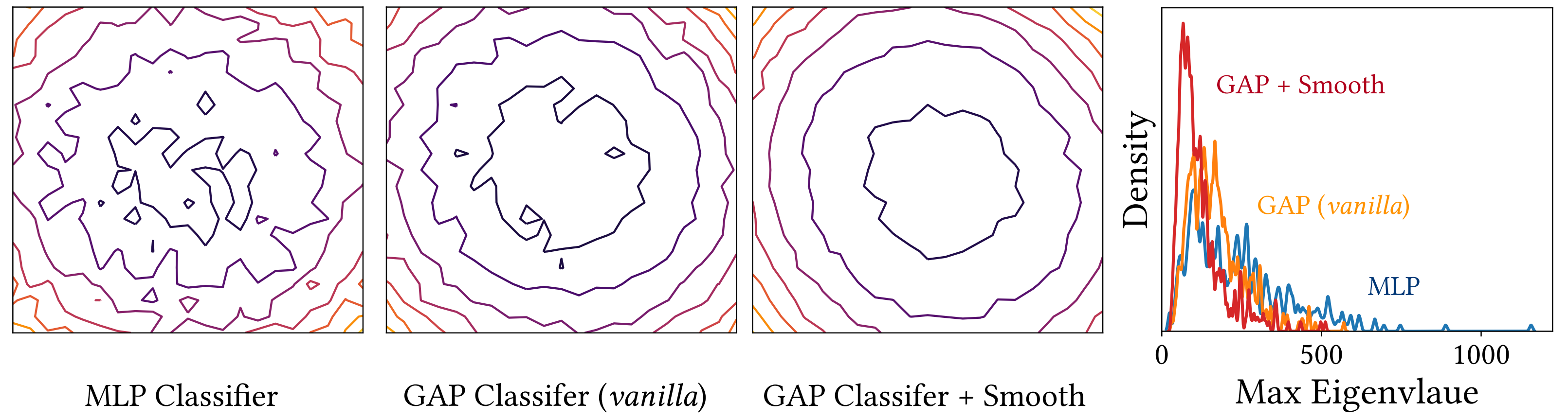
Figure 2. "MC dropout + spatial smoothing" is **25× faster** than canonical MC dropout with similar predictive performance. Spatial smoothing also improves deterministic NNs.

III. How Does Spatial Smoothing Improve NNs?



(a) Feature map of ResNet at the end of stage 1 (b) Fourier analysis

Figure 3. MC dropout adds high-frequency noises. Spatial smoothing **filters high-frequency signals** and stabilizes (denoises) feature maps.



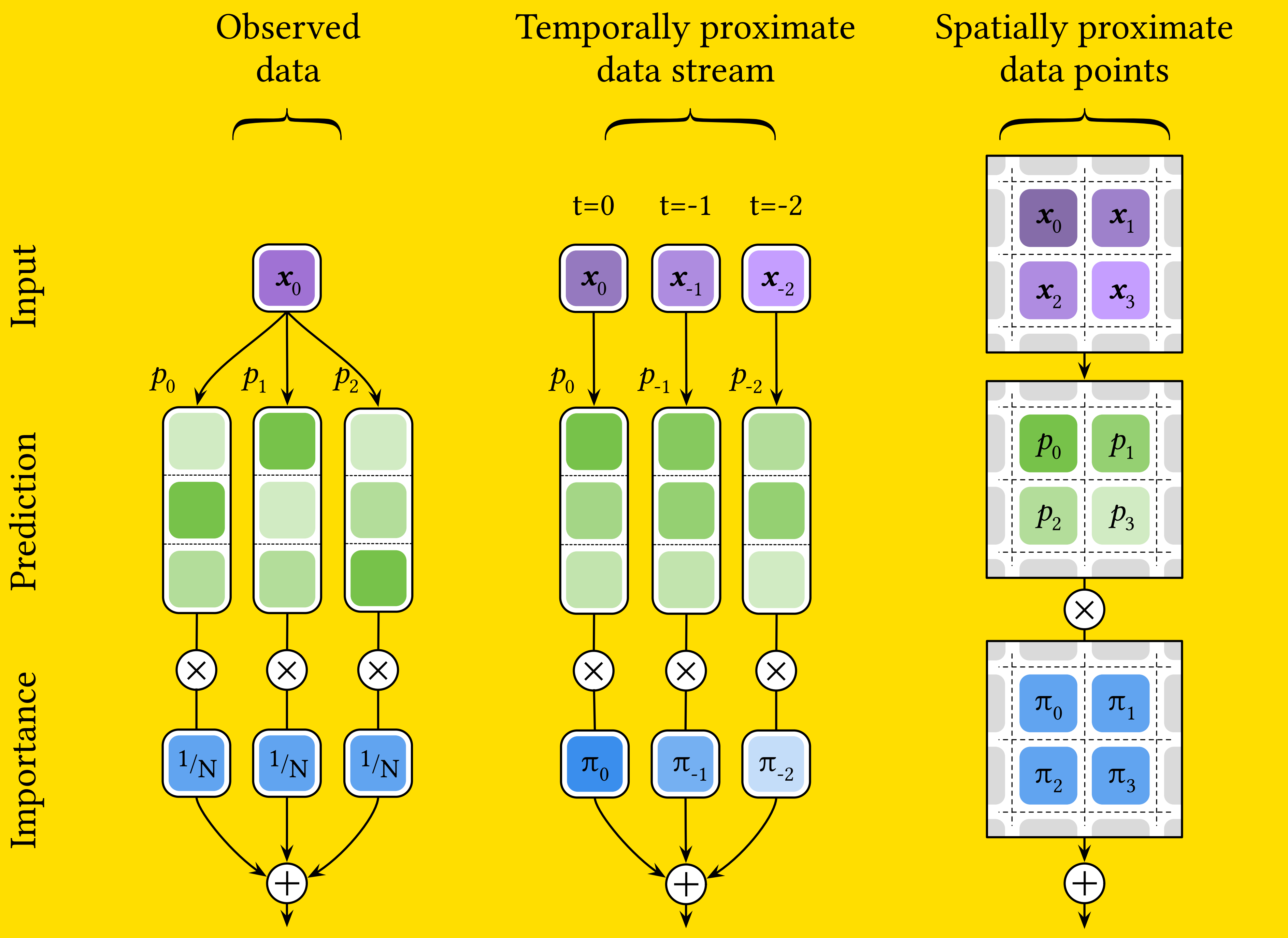
(a) Loss landscape visualization of ResNet (b) Hessian spectrum

Figure 4. Spatial smoothing helps NN optimization by **flattening loss landscapes**. In addition, GAP is an extreme case of spatial smoothing.

IV. How Can We Make Spatial Smoothing Trainable?

Self-attentions for computer vision, also known as Vision Transformers (ViTs), can be deemed as trainable importance-weighted ensembles of feature maps.

SCAN AND SEE A GITHUB

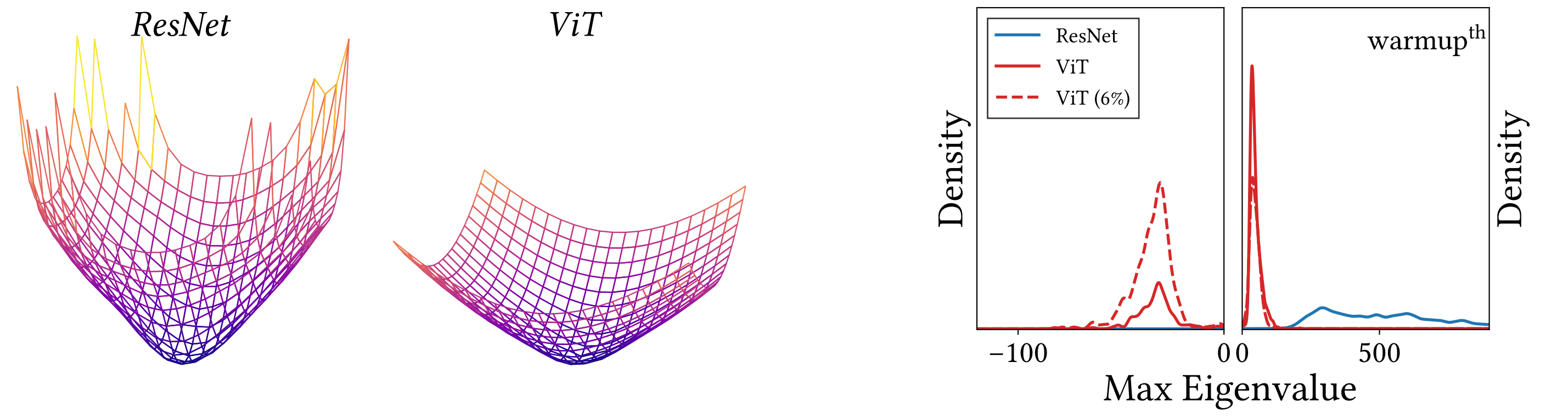


(a) Canonical ensemble (b) Temporal smoothing (c) Spatial smoothing (*ours*)

How Do Vision Transformers Work?

We provide an explanation of how ***self-attentions*** work by addressing them as ***a trainable spatial smoothing*** of feature maps, because the formulation suggests that self-attentions average feature map values with the positive importance-weights.

I. Self-Attentions Flatten Loss Landscapes



(a) Loss landscape visualization (b) Hessian max eigenvalue spectrum

Figure 5. Loss landscape visualization (*Left*) and Hessian max eigenvalue spectrum (*Right*) consistently show that **ViT has a flatter loss than ResNet**.

II. Self-Attentions Are Low-Pass Filters

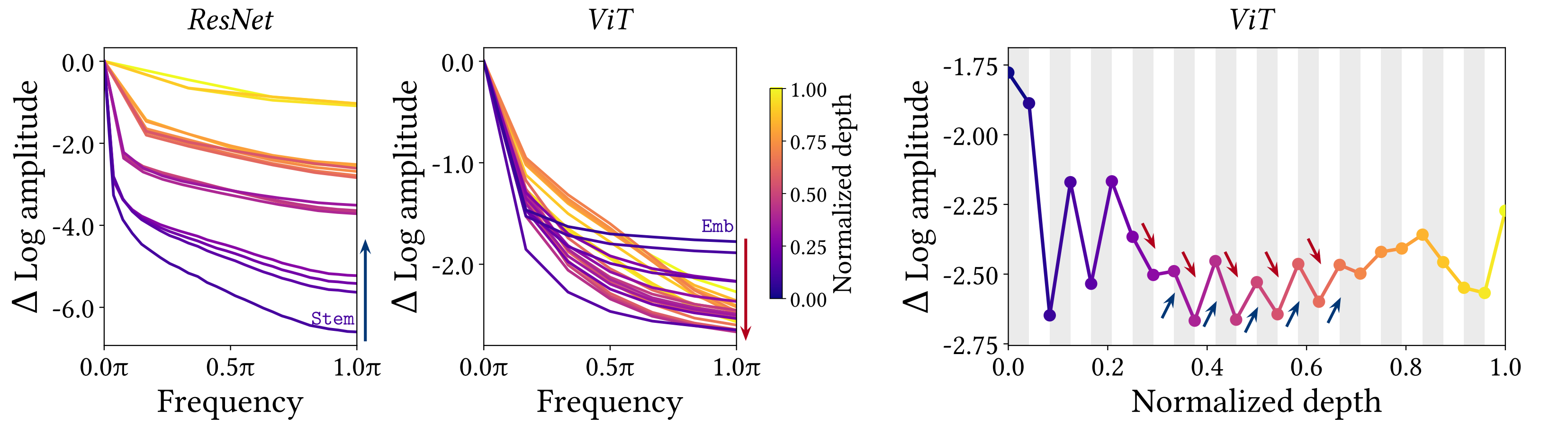
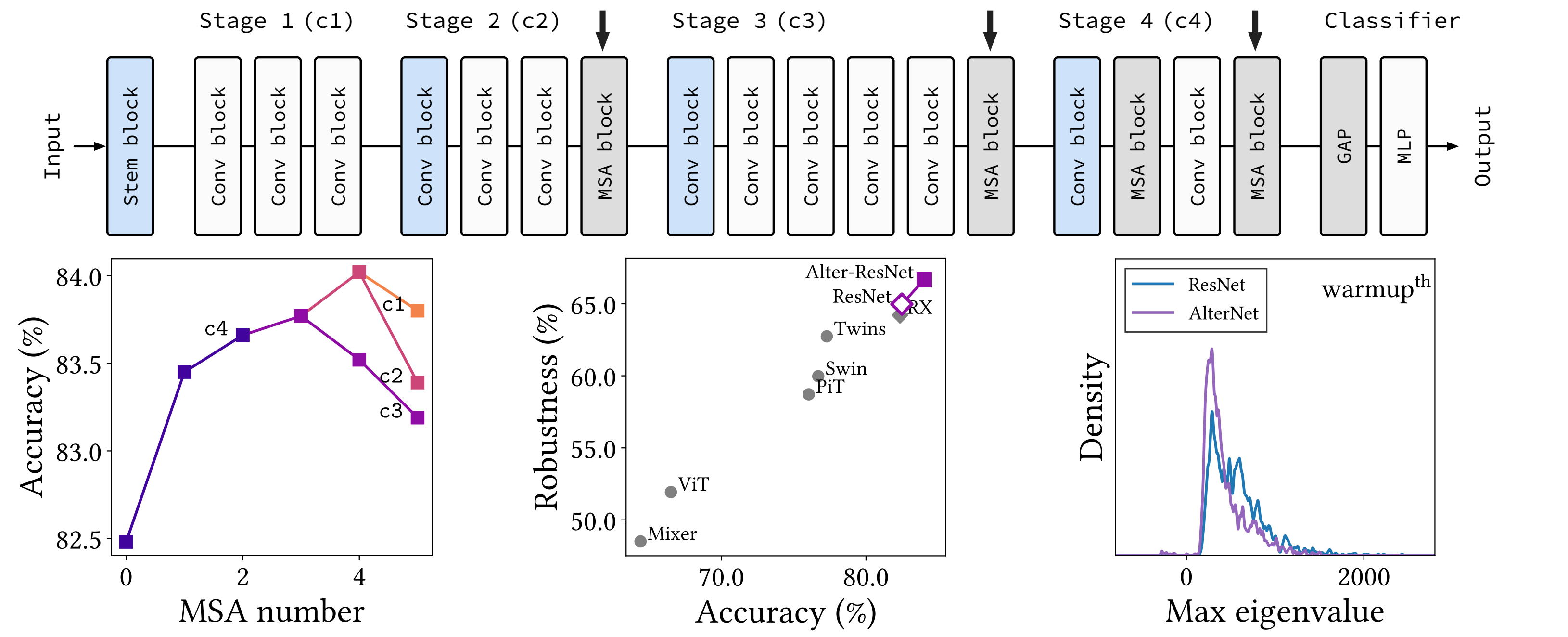


Figure 6. Relative log amplitudes of Fourier transformed feature map show that **ViT tend to reduces high-frequency signals, while ResNet amplify them**. *Right*: In ViT, MSAs (gray area) reduce the high-frequency (1.0π) component, and Conv/MLPs (white area) amplify it.

III. Self-Attentions at the End of a Stage Play a Key Role in Prediction



(b) Accuracy of AlterNet for MSA number (c) Accuracy and robustness in a small data regime (CIFAR-100) (d) Hessian max eigenvalue spectra

Figure 7. **Self-attentions at the end of a stage (not a model) and Convs at the beginning of a stage significantly improve the performance**. AlterNet, a model in which Conv blocks at the end of a stage are replaced with MSA blocks, outperforms CNNs even on CIFAR.

In summary, Self-attentions have a spatial smoothing inductive bias that complements convolutions.

	Self-Attention	Convolution
Loss Landscape	Flat but non-convex	Convex but sharp
Fourier Analysis	Low-pass filter (shape-biased)	High-pass filter (texture-biased)
Best Practice	The end of a stage	The beginning of a stage

SCAN AND SEE A GITHUB